Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 79-82 recibido 21-11-2017 revisado 07-01-2018 aceptado 06-02-2018

Detección de Patrones Psicolingüísticos para el Análisis de Lenguaje Subjetivo en Español

Psycholinguistic Patterns Detection for Analyzing the Subjective Language in Spanish

María del Pilar Salaz Zárate
Universidad de Murcia
Facultad de Informática Campus Espinardo
Espinardo, 30100, Murcia, España
mariapilar.salas@um.es

Resumen: Tesis doctoral titulada "Detección de patrones psicolingüísticos para el análisis de lenguaje subjetivo en español", defendida por María del Pilar Salas Zárate en la Universidad de Murcia y elaborada bajo la dirección de los doctores Rafael Valencia García (Universidad de Murcia) y Miguel Ángel Rodríguez García (Universidad King Abdulah). La defensa tuvo lugar el 23 de mayo de 2017 ante el tribunal formado por los doctores Jesualdo Tomás Fernández Breis (Presidente, Universidad de Murcia), Alejandro Rodríguez González (Secretario, Universidad Politécnica de Madrid) y José Antonio Miñarro Giménez (Vocal, Medical University of Graz) y la tesis obtuvo la mención Cum Laude y Doctora Internacional.
Palabras clave: Patrones psicolingüísticos, lenguaje subjetivo, minería de opiniones

Abstract: Ph.D. thesis entitled "Psycholinguistic patterns detection for analyzing the subjective language in Spanish" written by María del Pilar Salas Zárate at the University of Murcia under the supervision of the Ph.D. Rafael Valencia García (University of Murcia) and Ph.D. Miguel Ángel Rodríguez García (University). The viva voice was held on the 23rd may 2017 and the members of the commission were the Ph.D. Jesualdo Tomás Fernández Breis (President, University of Murcia), Ph.D. Alejandro Rodríguez González (Secretary, Polytechnic University of Madrid) and Ph.D. José Antonio Miñarro Giménez (Vocal, University of Graz) and the thesis obtained the mention Cum Laude and International Doctor.
Keywords: Psycholinguistic patterns, subjective language, opinion mining

1 Introducción

Las opiniones son una parte importante en las decisiones del ser humano, cuando una persona desea tomar una decisión se basa en los comentarios de otras personas, por ejemplo, para comprar un producto, seleccionar un destino turístico, incluso para votar por un partido político. Con el surgimiento de la Web 2.0, ya no sólo se dependía de las opiniones de familiares o amigos, sino que se podía acceder a una gran cantidad de información en la Web provista por otros usuarios. Por lo que, actualmente, las personas visitan blogs, foros de discusión o redes sociales con el objetivo de obtener las experiencias de otros usuarios antes de tomar una decisión.

La lingüística es una de las áreas que se ha enfocado en el estudio de la opinión, o mejor dicho del lenguaje subjetivo. Este tipo de lenguaje se emplea para expresar estados personales en el contexto de una conversación o un texto (Wiebe, Wilson, Bruce, Bell, y Martin, 2004; Martínez-Cámara, 2016). Por otro lado, el análisis de sentimientos, también conocido como minería de opiniones, se ha convertido en un tema muy popular que tiene como objetivo el procesar opiniones públicas disponibles en la Web a través de técnicas de procesamiento de lenguaje natural. En este contexto, diferentes propuestas basadas en aprendizaje automático y orientación semántica han surgido en los últimos años. Estos trabajos
abordan problemas tales como análisis y construcción de lexicones de sentimientos,

ISSN 1135-5948. DOI 10.26342/2018-60-10

evaluación y clasificación de mensajes de Twitter, negación, por mencionar algunos. Otros trabajos se centran en analizar las opiniones en diferentes niveles, a saber, documento, sentencia y aspectos.

A pesar de los esfuerzos llevados a cabo en el análisis de sentimientos existen diversas características psicológicas y lingüísticas que no han sido abordadas. Por lo tanto, la clasificación automática de opiniones requiere un esfuerzo multidisciplinario, donde la lingüística y el procesamiento del lenguaje natural juegan un rol importante. Gracias a estas disciplinas es posible entender el lenguaje humano, clasificar las opiniones y resumir los sentimientos expresados acerca de un producto, servicio o cualquier otro aspecto.

El lenguaje figurado tal como la ironía, el sarcasmo y la sátira juega un papel muy importante en los sistemas de análisis de sentimientos. El doble sentido expresado en una opinión o comentario a través de este lenguaje puede invertir la polaridad de la opinión. Aunque, el lenguaje figurado ha sido ampliamente estudiado por diversas áreas como la lingüística, solo pocos estudios se han enfocado en la detección automática.

Por otro lado, es importante mencionar que pocos trabajos para el análisis de sentimientos, y lenguaje figurado, se han enfocado en el idioma español, quizá debido a la carencia de recursos lingüísticos en ese idioma. Sin embargo, el estudio del español es de lo más importante ya que es uno de los idiomas más utilizados en internet.

Las razones expuestas en los párrafos anteriores han sido la principal motivación para la realización de esta tesis doctoral. Por lo que se propone un método para la detección de patrones psicolingüísticos para el análisis de sentimientos y la detección de la sátira en español. Este método permite, a través de un enfoque automático supervisado, clasificar textos como positivo, negativo, neutro, muy positivo o muy negativo y como satíricos y no satíricos.

2 Objetivos

El objetivo principal de esta tesis doctoral es la detección de patrones psicolingüísticos para el análisis de lenguaje subjetivo en español. Específicamente, se propone el desarrollo de un método para el análisis de sentimientos y la detección de textos satíricos y no satíricos. Por lo tanto, los siguientes puntos fueron abordados en este trabajo.

Determinar qué tan relevantes son las características psicolingüísticas en la clasificación de sentimientos.

Determinar qué tan relevantes son las características psicolingüísticas en la clasificación de textos satíricos.

Identificar cuáles son las características más relevantes para el análisis de sentimientos.

Identificar cuáles son las características más relevantes para la detección de la sátira

3 Estructura de la tesis

La tesis doctoral se divide en cinco capítulos que exponen el estudio que se realizó. A continuación, se describe brevemente el contenido de cada uno de estos capítulos.

Capítulo 1. Este capítulo provee una breve introducción al trabajo de investigación, incluyendo la problemática a abordar.

Capítulo 2. Este apartado consiste en un detallado estudio de la bibliografía relacionada con las tecnologías base para el desarrollo del método propuesto. El estudio inicia con una introducción al lenguaje subjetivo. Posteriormente, se presenta el campo del procesamiento del lenguaje natural, así como los diferentes niveles de procesamiento. Después, se describe el campo del análisis de sentimientos y se proporcionan las definiciones más utilizadas por la comunidad investigadora. Además, se presenta su evolución histórica desde sus inicios en el siglo XX hasta la fecha. Asimismo, se presentan los diferentes niveles de análisis de opiniones, así como los dos principales enfoques en los cuales se basan la mayoría de los estudios, a saber, orientación semántica y aprendizaje automático. En el penúltimo apartado, se provee una introducción al lenguaje figurado, específicamente la ironía, el sarcasmo y la sátira. Finalmente, se presenta la importancia de las características psicolingüísticas en el lenguaje humano, y se introduce la herramienta LIWC, la cual permite obtener variables psicolingüísticas desde un texto escrito.

Capítulo 3. Este capítulo describe el método para el análisis de sentimientos y la detección de la sátira propuesto en este trabajo de investigación.

Capítulo 4. Esta sección se centra en la validación del método propuesto para el análisis de sentimientos y detección de la sátira. Este

capítulo se divide en dos apartados. En el primer apartado se presentan y discuten los resultados del análisis de sentimientos en dos dominios, a saber, películas y turismo. En el segundo apartado se presentan los resultados y discusión para la detección de la sátira, el cual fue validado en el dominio de noticias.

Capítulo 5. Finalmente, en este capítulo presentan las conclusiones obtenidas del trabajo de investigación y las posibles vías futuras.

4 Contribuciones

Las principales aportaciones de esta tesis se resumen a continuación.

Desarrollo de un método para la clasificación de sentimientos y detección de la sátira. Este método permite clasificar opiniones como positivas, negativas, neutras, muy positivas y muy negativas y tweets como satíricos y no satíricos. El método puede ser adaptado a diversos problemas de clasificación de textos e idiomas. Sin embargo, este requiere un corpus etiquetado como entrada.

Proceso para el preprocesamiento de tweets en español: La normalización de textos extraídos de redes sociales tal como Twitter suele ser más difícil debido a que los usuarios suelen abreviar palabras y usar jerga debido a la limitación de 140 caracteres que tienen los tweets. Actualmente, existen pocas herramientas del procesamiento del lenguaje natural que permiten normalizar estos textos en español. Para ello, nosotros definimos un proceso que permite normalizar los tweets para procesarlos posteriormente como un texto normal. El proceso consiste en tres principales pasos: 1) tokenización del texto y detección de entidades tales como URLs, menciones y etiquetas; 2) eliminar los elementos detectados en el paso 1 con excepción de etiquetas donde sólo es eliminado el "#"; y 3) extensión de abreviaturas y corrección de ortografía. Este proceso es de suma importancia, ya que actualmente Twitter está siendo un foco de investigación debido a la gran cantidad de información subjetiva contenida en estas redes sociales, la cual está constituida principalmente por opiniones.

Desarrollo de un corpus en el dominio del turismo. Los corpora son un recurso importante en el análisis de sentimientos. Por un lado, los métodos basados en un enfoque de aprendizaje automático requieren de corpus etiquetados con el objetivo de entrenar algoritmos de clasificación. Por otro lado, estos corpus sirven como base para la evaluación de sistemas de análisis de sentimientos. El desarrollo de un corpus requiere esfuerzo y tiempo debido a que el etiquetado se realiza manualmente con el objetivo de obtener un corpus de calidad. Sin embargo, hoy en día existen pocos corpus disponibles en español en la comunidad investigadora, es por ello, que el corpus obtenido en este trabajo de tesis supone una gran aportación.

Desarrollo de un corpus de tweets satíricos. Este corpus consiste en un conjunto de tweets etiquetados como satíricos y no satíricos extraídos desde diversas cuentas de Twitter. Actualmente existen algunos corpus del lenguaje figurado como ironía y sarcasmo. Sin embargo, hay una carencia de corpus con información satírica, sobre todo en español. Este corpus además de ser en este idioma está dividido en sátira mexicana y satírica española.

Detección de características psicolingüísticas para el análisis de sentimientos. Otra aportación relevante de esta tesis se centra en la identificación y extracción de características psicolingüísticas que son más discriminantes para el análisis de sentimientos y detección de la sátira.

5 Líneas futuras

Con respecto a investigación futura, se proveen varios aspectos que no han sido considerados como parte de esta tesis. Sin embargo, son considerados como líneas de investigación futuras a explorar. A continuación, se detalla cada uno de estos aspectos.

Integrar técnicas que permitan proveer un mejor soporte del proceso de normalización ante casos como tweets. La normalización de textos como tweets es una tarea muy difícil debido a que normalmente son textos con palabras abreviadas y con faltas de ortografía. En este trabajo de tesis, se propone un proceso para su normalización. Sin embargo, tiene una limitación en cuanto al procesamiento de etiquetas en inglés hashtags. Una etiqueta puede contener múltiples palabras juntas. Por lo que considerar técnicas tales como la presentada en (Bejcek, Stranák, y Pecina, 2013), permitirá abordar este problema.

Aplicación del método a diversos dominios. El método de análisis de sentimiento desarrollado en esta tesis ha sido favorablemente aplicado en los dominios

turístico y de películas. Por lo que la aplicación a otros dominios sería otra de las líneas de investigación a explorar como posible línea futura para tener en cuenta. Sin embargo, como se mencionó anteriormente, se requiere de un corpus del dominio etiquetado. Es por ello, que se propone el desarrollo de nuevos corpus en diversos dominios. Un área de especial interés es el dominio médico, el cual ha sido poco explorado. Sin embargo, las opiniones pueden ser de gran interés entre pacientes sobre todo cuando padecen de enfermedades que requieren de autogestión como la diabetes, asma, cáncer, hipertensión, etc. Por otro lado, en cuanto al lenguaje figurado, este trabajo está enfocado en la detección de la sátira, por lo que la creación de nuevos corpus para la ironía, sarcasmo y sátira en diversos dominios permitiría el desarrollo y evaluación de nuevos sistemas de análisis de sentimiento.

Aplicación del método en otros idiomas. La aplicación del método se ha enfocado en el idioma español, por lo que como trabajo a futuro se aplicará este método a otros idiomas tales como inglés, francés, árabe y a diversas variedades de español, como el que se habla en Argentina, Uruguay, Venezuela, etc. Esto permitirá determinar si los patrones psicolingüísticos detectados en esta tesis pueden contribuir también a la detección del análisis de sentimientos y sátira en diferentes idiomas y culturas.

Detección de patrones psicolingüísticos para el sarcasmo y la ironía. El procedimiento para detectar patrones psicolingüísticos únicamente ha sido diseñado para la sátira. Por lo que sería muy interesante también detectar patrones psicolingüísticos para el sarcasmo e ironía. Además, esto permitiría determinar el nivel de similitud entre estos tipos de lenguaje figurado, es decir, determinar qué categorías psicolingüísticas comparten.

Integración del sistema de detección del lenguaje figurado en el análisis de sentimientos. Los sistemas presentados en este trabajo de tesis doctoral son independientes. Por lo que la incorporación de un módulo que permita detectar la ironía, el sarcasmo y la sátira, así como otros tipos de lenguaje figurado como el humor en el sistema de análisis de sentimientos, permitirá no sólo detectar la polaridad de la opinión, sino también detectar si el texto es literal, irónico, sarcástico o satírico.

Contribuir al enriquecimiento de LIWC en español. La extracción de características depende en gran medida del diccionario de LIWC. Sin embargo, este diccionario carece de algunas palabras del español como verbos y de una gran variedad de palabras utilizadas en diversos países como Venezuela, Colombia, Ecuador, etc. El diccionario de LIWC puede ser enriquecido con otras palabras, lo cual se traduciría en una mejor extracción de características y, por tanto, una mejor precisión del sistema.

Agradecimientos

María del Pilar Salas Zárate es apoyada por la Comisión Nacional de Ciencia y Tecnología (CONACyT) y la Secretaria de Educación Pública (SEP).

Bibliografía

Bejcek, E., P. Stranák, y P. Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. En *Proceedings of the 9th Workshop on multiword expressions,* páginas 106-115.

Martínez-Cámara, E. 2015. Análisis de opinions en Español (Tesis de doctorado). Universidad de Jaén. Departamento de Informática. Obtenido de http://rua.ua.es/dspace/handle/10045/53569

Wiebe, J., T. Wilson, R. Bruce, M. Bell, y M. Martin. 2004. Learning Subjective Language. *Computational Linguistics.*, 30(3), 277-308.

Salas-Zárate, M. d. P., M. A. Paredes-Valverde, M. A. Rodriguez-Garcia, R. Valencia-García, y G. Alor-Hernández. 2017. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128, 20-33.

Salas-Zárate, M. d. P., R. Valencia-García, A. Ruiz-Martínez, y R. Colomo-Palacios. 2017. Feature-based opinion mining in financial news: an ontology-driven approach. *Journal of Information Science*, 43(4), 458-479.

Salas-Zarate, M. d. P., M. A. Paredes-Valverde, J. Limon, D. A. Tlapa, & Y. A. Báez. 2016. Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods. *Journal of Universal Computer Science*, 22(5), 691-708.